Spatio-Temporal Analysis of Wikipedia Metadata and the STiki Anti-Vandalism Tool

Andrew G. West • Sampath Kannan • Insup Lee

Department of Computer and Information Science -- University of Pennsylvania -- Philadelphia, PA, USA

BIG IDEA

Spatio-temporal properties of edit metadata (editor, article, timestamp, and revision comment) can be leveraged to detect Wikipedia vandalism comparably to NLP based methods:

- Simple features (*i.e.*, time-of-day), in addition to historical *reputations* for editors, articles, and spatial groupings thereof are used.
- Such features have language-independence, efficiency, and robustness not found in traditional detection mechanisms (*i.e.*, NLP).
- STiki [1], is a real-time, on-Wikipedia tool utilizing the technique, already shown feasible off-line in our prior work [3].

EDIT LABELING: ROLLBACK

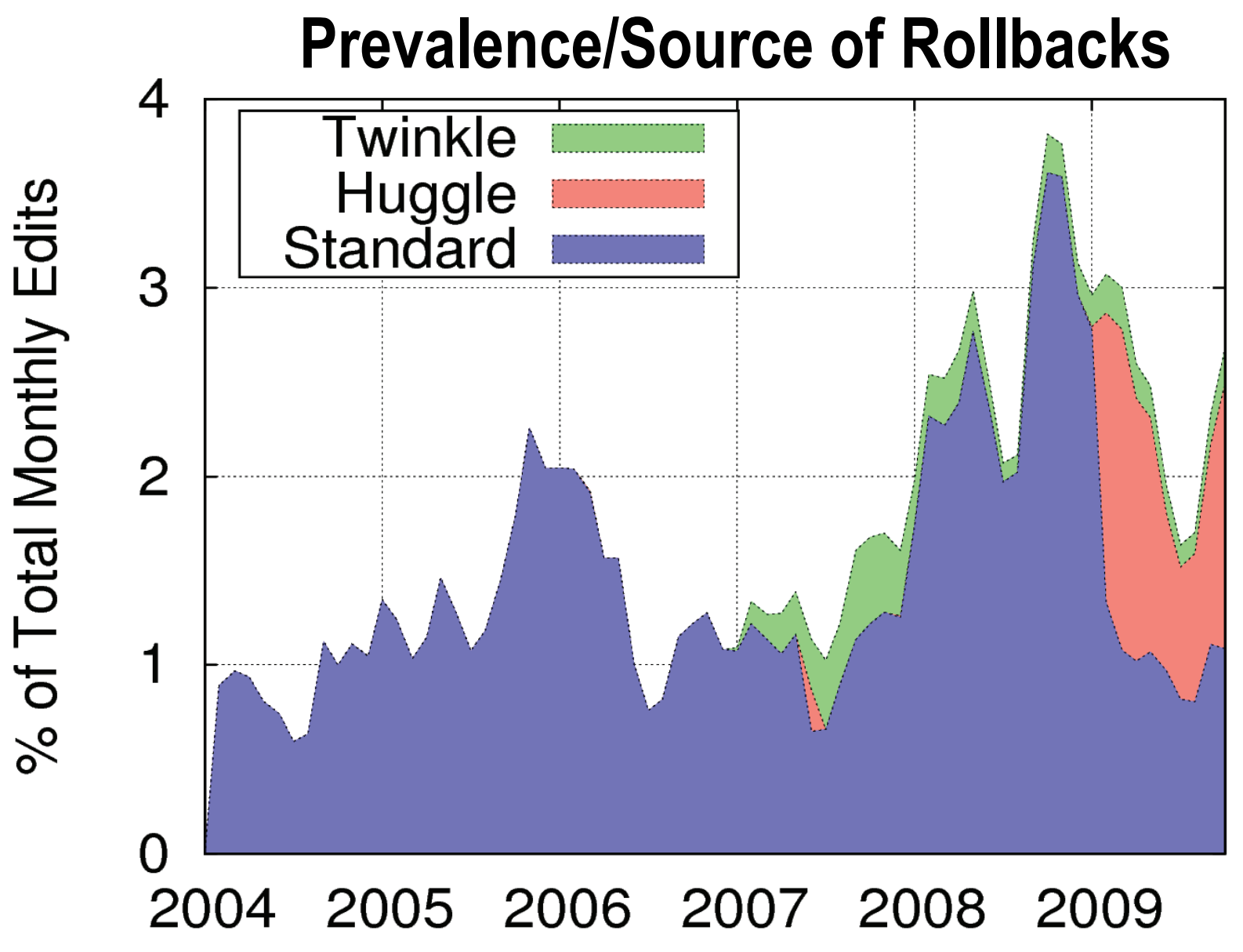
Need to label edits exhibiting vandalism (ex-post facto) to:

1. Show *features* effective (and eventually to train over them)
2. Form basis of historical reputations (vandalism = misbehavior)

ROLLBACK

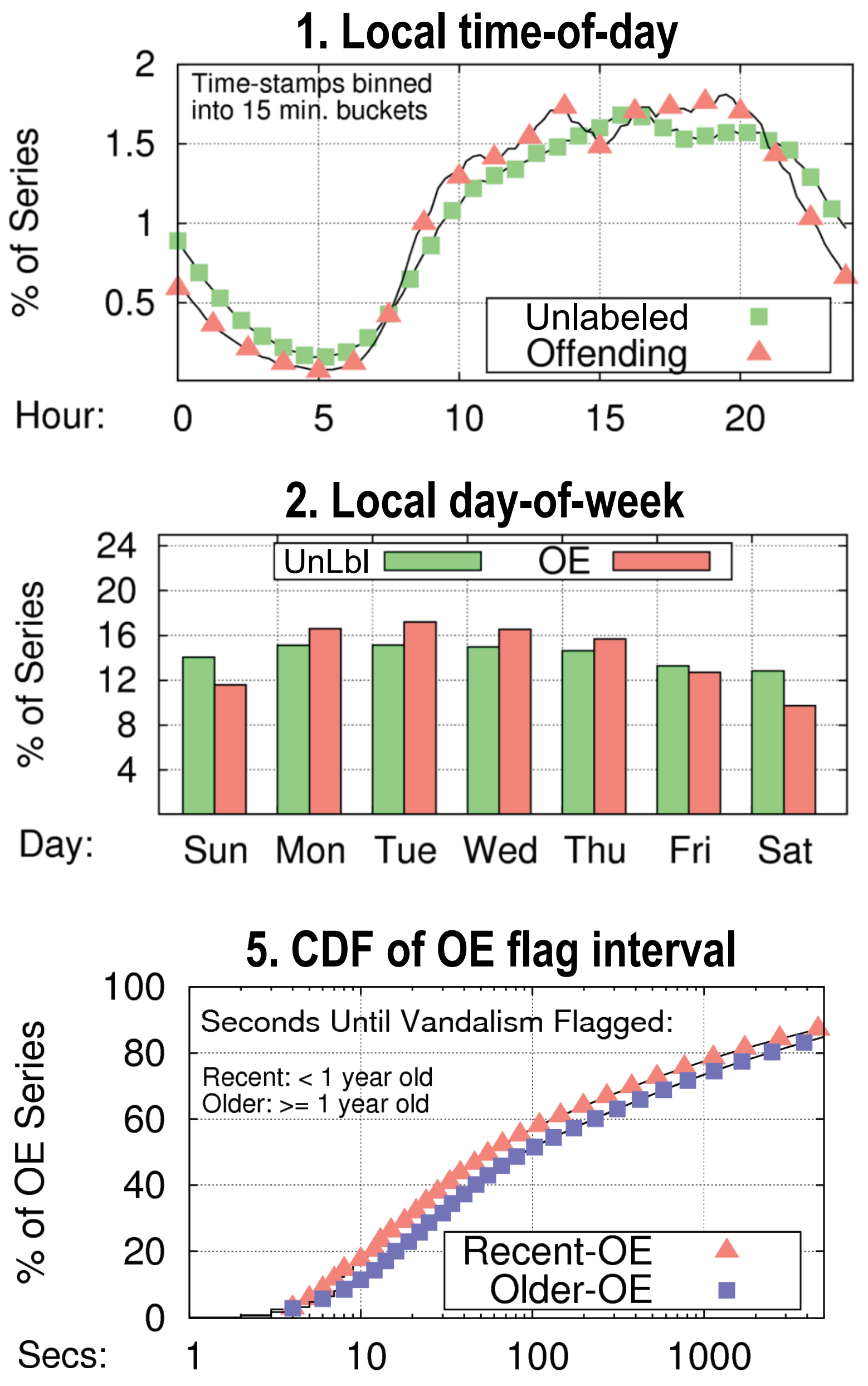
An administrative form of [undo]:

- Revisions undone are Offending Edits (OEs), likely vandalism
- Autonomously parse-able
- Trusted feedback (admins)
- Vandalism defined case-by-case



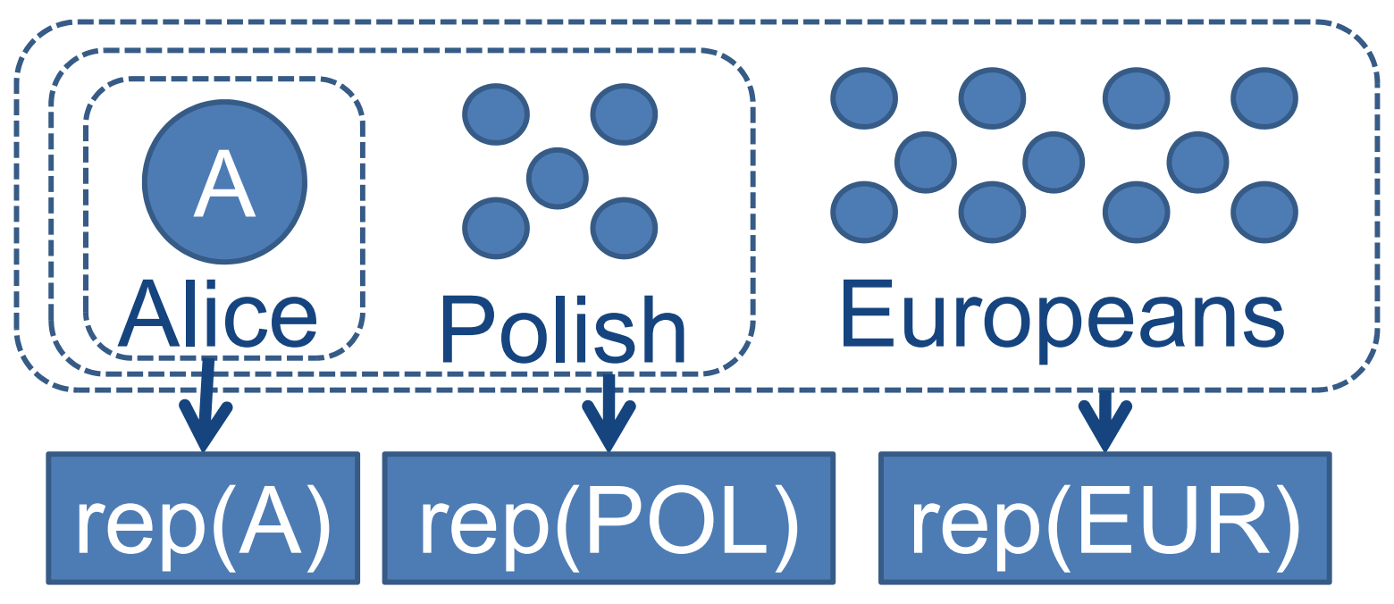
SIMPLE SPATIO-TEMPORAL FEATURES

#	FEATURE
1	**Edit time-of-day**: (see right)
2	**Edit day-of-week**: (see right)
3	**Time-since article edited**: Frequently edited pages are vandalism targets (visibility)
4	**Time-since editor reg.**: Long-time editors are rarely problematic (Sybil attack)
5	**Time-since last user OE**: Good editors rarely vandalize (+OEs flagged quickly; see left)
6	**Revision comment length**: Vandals leave shorter comments (lazy + bandwidth)



AGGREGATE FEATURES (REPUTATIONS)

IDEA: Use entity-specific reputation; augment with spatial reputations [2], which will have more historical data.



#	FEATURE
7	**Article reputation**
8	**Category reputation** Spatial grouping over articles
9	**Editor reputation**
10	**Country reputation** Spatial grouping over editors

The reputation function:

- Summation over time-decayed feedback (vandalism via rollback)
- Spatial reputation's are normalized by the group size

RANK	COUNTRY	%-OEs
1	Italy	2.85%
2	France	3.46%
13	United States	11.63%
14	Australia	12.08%

ARTICLE*	#OEs
Wikipedia	5589
United States	2161
World War II	1886

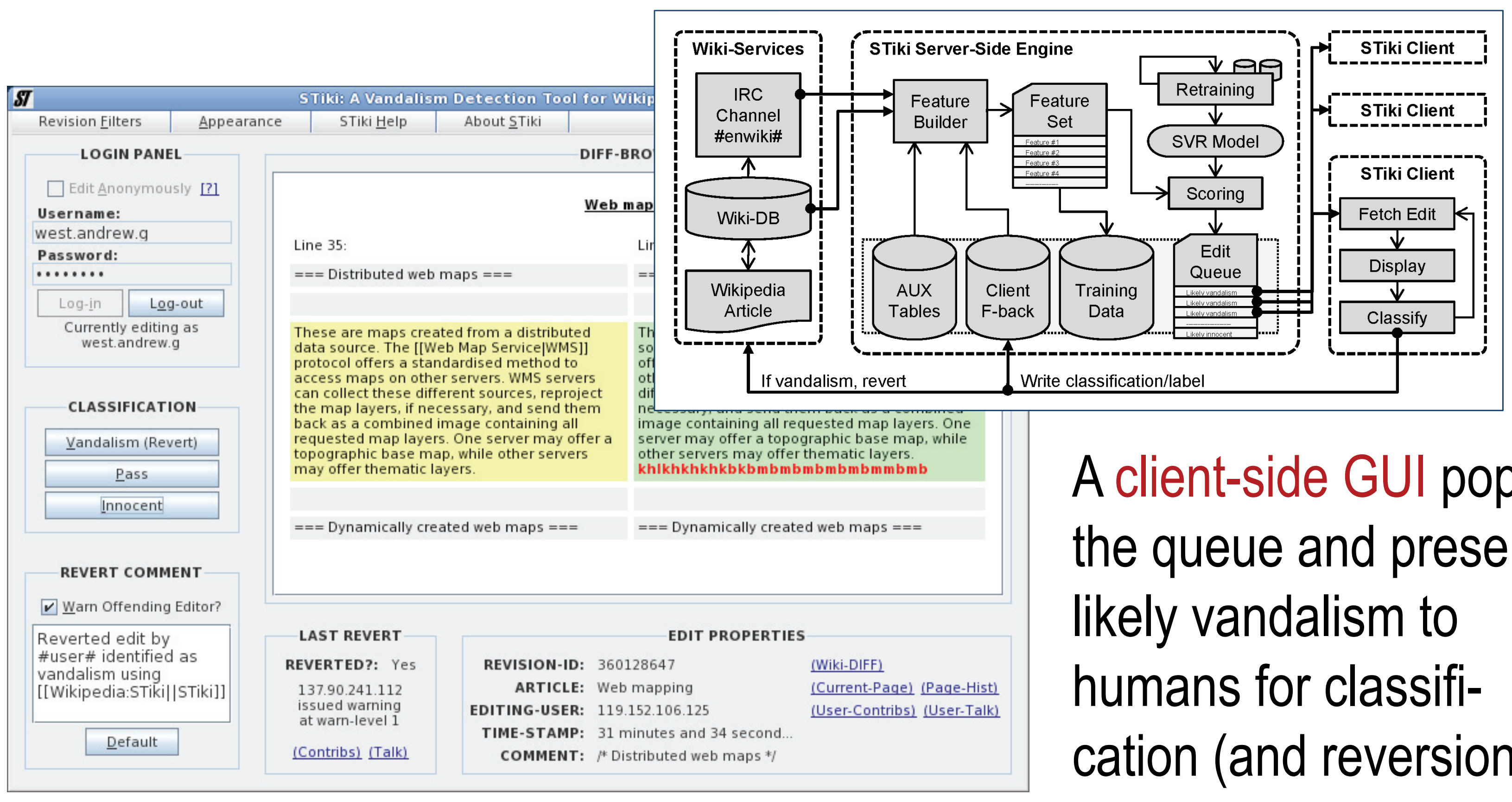
* List sanitized for appropriateness

CATEGORY (w//100+ pgs)	PGs	OEs/PG
World Music Award Winners	125	162.27
Characters of Les Miserables	135	146.88
Former British Colonies	145	141.51

Vandalism is clustered non-uniformly throughout article and editor space, making membership in such groupings behavior predictive.

THE STiki TOOL

STiki [1] leverages these features in real-time. The server-side engine calculates a real valued *vandalism score* (via machine-learning) for all edits, which is the insertion priority into the *edit queue*.



STiki: A Vandalism Detection Tool for Wikip

Revision Filters | Appearance | STiki Help | About STiki

LOGIN PANEL

☐ Edit Anonymously [?]

Username: west.andrew.g

Password: ••••••••

Log-in | Log-out

Currently editing as west.andrew.g

CLASSIFICATION

Vandalism (Revert) | Pass | Innocent

REVERT COMMENT

☑ Warn Offending Editor?

Reverted edit by #user# identified as vandalism using [[Wikipedia:STiki||STiki]]

Default

DIFF-BRO

Web map

Line 35:

=== Distributed web maps ===

These are maps created from a distributed data source. The [[Web Map Service|WMS]] protocol offers a standardised method to access maps on other servers. WMS servers can collect these different sources, reproject the map layers, if necessary, and send them back as a combined image containing all requested map layers. One server may offer a topographic base map, while other servers may offer thematic layers.

image containing all requested map layers. One server may offer a topographic base map, while other servers may offer thematic layers.
khlkhkhkhkbkbmbmbmbmbmbmmbmb

=== Dynamically created web maps ===

=== Dynamically created web maps ===

LAST REVERT

REVERTED?: Yes

137.90.241.112 issued warning at warn-level 1

(Contribs) (Talk)

EDIT PROPERTIES

REVISION-ID: 360128647 (Wiki-DIFF)

ARTICLE: Web mapping (Current-Page) (Page-Hist)

EDITING-USER: 119.152.106.125 (User-Contribs) (User-Talk)

TIME-STAMP: 31 minutes and 34 second...

COMMENT: /* Distributed web maps */

A client-side GUI pops the queue and presents likely vandalism to humans for classification (and reversion).

An edit is also de-queued if a more recent edit is made on the same article.

STiki PERFORMANCE & FUTURE

Performance metric: *hit-rate* (% of displayed edits that are vandalism):

- Offline-analysis [3] shows hit-rate should be 50%+
- In fact, ≈10% due to competing tools/bots (often autonomous)

Successes and alternative uses:

- STiki has reverted over 2000 instances of vandalism on *en-wiki*.
- Combats embedded vandalism well. Median age of vandalism reverted by STiki is 4.25 hours, nearly 200× of conventional reverts.
- May be best suited for less-patrolled Wikis (*e.g.* foreign lang. eds.)

Future improvements:

- Include lightweight NLP features (a general-purpose tool)
- Alternative detection (link spam? more ST-features?)

REFERENCES & ACKNOWLEDGEMENTS

[1]: A. G. West. STiki: A vandalism detection tool for Wikipedia. *http://en.wikipedia.org/wiki/Wikipedia:STiki*, 2010. Software.

[2]: A. G. West, A. J. Aviv, J. Chang, and I. Lee. Mitigating spam using spatio-temporal reputation. *Technical Report MS-CIS-10-04, University of Pennsylvania*, Feb. 2010.

[3]: A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC `10: Proc. of the 3rd European Workshop on System Security*, pages 22-28, Paris, France, Apr. 2010.

This research was supported in part by ONR MURI N00014-07-1-0907. POC: Insup Lee, lee@cis.upenn.edu